INTRODUCTION
A L'ANALYSE DES DONNEES
ET
A LA REGRESSION

La statistique est une branche des mathématiques appliquées considérée en France par les mathématiciens comme mineure. Elle est marginalisée au plan de la recherche scientifique et souvent au plan de l'enseignement, au point que dans certaines facultés des sciences, elle n'est pas enseignée dans le cursus des étudiants en mathématiques (c'est d'ailleurs très regrettable quand on sait que certains des étudiants devront l'enseigner ultérieurement en collège ou en lycée). Cette situation particulière au sein de l'Université française s'explique par la nature même de cette discipline, dont les méthodes sont développées par des mathématiciens pour être appliquées par des utilisateurs qui n'en sont pas.

En fait, la statistique appliquée est une approche du réel fondée sur la répétition d'expériences et d'observations : cette démarche scientifique est utilisée dans de nombreuses sciences appliquées ou humaines telles que la psychologie (certains psychologues américains, comme Spearman, sont d'ailleurs à l'origine de l'analyse factorielle), la sociologie, la médecine .... On peut distinguer deux démarches différentes : la première, connue sous le nom de statistique descriptive, se limite à l'analyse des données recueillies, et la seconde , la statistique inférentielle, évalue la part du hasard dans les observations effectuées pour en extraire autant que possible des propriétés générales.

Cet ouvrage s'inscrit surtout dans la démarche de la statistique descriptive. Il s'inspire, dans ses grandes lignes, d'enseignements effectués dans des établissements universitaires de niveaux et d'orientations variés : DEUG Math. Physique, I.U.T. Informatique, Techniques de Commercialisation, maîtrise d'Ingénierie Mathématique, écoles de Commerce et de Gestion. Il présente les méthodes principales de l'analyse des données et la régression linéaire multiple dont l'importance s'est fortement accrue depuis que les problèmes numériques posés par ces méthodes ont pu être résolus à l'aide d'ordinateurs et de logiciels.

Le niveau mathématique des étudiants auxquels il s'adresse n'est pas nécessairement élevé : son objectif initial est essentiellement de permettre au lecteur de choisir les méthodes adaptées au traitement statistique de leurs données, de les mettre en oeuvre et d'en interpréter les résultats correctement. L'ouvrage ne contient donc aucune démonstration, au sens mathématique du terme. Mais le lecteur est supposé connaître les éléments classiques de la statistique descriptive : moyennes, variances, écarts-types, histogrammes, coefficients de corrélation linéaire. Pour appliquer certaines méthodes, des notions élémentaires sur les tests statistiques sont nécessaires mais rappelées dans l'ouvrage.

Les méthodes présentées sont les plus courantes : analyse en composantes principales, analyse des correspondances, analyse discriminante, régression linéaire multiple, analyse factorielle des correspondances multiples, analyse factorielle d'un tableau de distances.

L'analyse en composantes principales (chapitre I) concerne des données quantitatives : elle donne une description des unités statistiques et des variables observées fondée sur l'étude des coefficients de corrélation linéaire. Des exemples numériques détaillés expliquent comment appliquer les formules.

Les tableaux de données établis à partir de données qualitatives peuvent être étudiés par l'analyse des correspondances (chapitre II), qui généralise en quelque sorte le test d'indépendance du $\chi 2$ de Pearson sur les tableaux de contingence et détaille la liaison entre les deux variables observées. Nous la présentons comme une double analyse en composantes principales.

L'analyse discriminante (chapitre III) utilise une part plus importante de la statistique inférentielle. Son objectif est de classer des unités statistiques dans des groupes donnés a priori, en fonction des valeurs qu'elles prennent sur des variables quantitatives. La démarche utilisée est habituelle : nous recherchons des variables quantitatives permettant de distinguer le mieux possible les groupes donnés et définissons des règles de décision.

Le chapitre IV est consacré à la régression linéaire multiple. Après avoir expliqué la démarche classique, nous critiquons les résultats d'un exemple numérique et en déduisons des méthodes particulières plus efficaces que la méthode classique. Ce chapitre utilise les résultats élémentaires de la statistique inférentielle, en particulier les tests de Student et de Fisher Snedecor.

On aborde, dans le chapitre V, l'analyse factorielle des correspondances multiples, dont l'objectif est l'étude simultanée des liaisons entre plusieurs variables qualitatives observées sur un même ensemble d'individus. Cette généralisation de l'analyse des correspondances est facile à mettre en oeuvre et très souvent utilisée dans les sciences humaines, mais elle donne des résultats parfois difficiles à interpréter.

Nous avons placé en chapitre VI l'analyse factorielle d'un tableau de dissimilarités et une introduction à la classification automatique. L'analyse d'un tableau de dissimilarités présente un intérêt pédagogique évident puisqu'elle introduit une notion fondamentale de l'analyse des données : les données analysées ne sont pas toujours obtenues par un tirage aléatoire dans une population statistique, et le modèle probabiliste n'est guère utilisé. En d'autres termes, les méthodes de description regroupées au sein de l'analyse des données ne nécessitent en général aucune hypothèse théorique sur les distributions de probabilité des observations. Cette analyse donne en outre un codage numérique des données souvent utilisé pour classer les observations à l'aide d'algorithmes : il s'agit de la classification automatique dont nous donnons un rapide aperçu.

Les chapitres sont ainsi chacun consacrés à une méthode statistique particulière. Pour faciliter l'utilisation du manuel, ils sont rédigés indépendamment les uns des autres et on trouvera donc parfois des redites. En général, l'exposé méthodologique est constitué de trois parties :

- la première indique la finalité de la méthode et la nature des données étudiées :
- la seconde donne les principes fondamentaux de la méthode :
- la troisième contient des compléments, dont l'intérêt consiste la plupart du temps à faciliter l'interprétation des résultats.

Un exemple numérique est traité entièrement tout au long de ces trois parties, et permet au lecteur d'effectuer les calculs proposés à l'aide d'une simple calculatrice. Tous les résultats numériques sont donnés pour qu'il puisse vérifier ses résultats. Il est clair que cette démarche est limitée à ces cas d'école, et qu'il n'est pas question, sauf exception, d'effectuer ce genre de calcul dans la pratique.

L'exposé méthodologique est suivi d'une ou plusieurs études de cas montrant les propriétés de la méthode statistique. Certaines études de cas sont corrigées : pour les autres, nous avons choisi de donner les résultats numériques (que le lecteur devrait lui-même établir par ordinateur) et de guider l'étudiant par des questions précises. Elles traitent de sujets divers, et les données sont réelles ou simulées suivant le but recherché, qui peut être de montrer l'efficacité de la méthode statistique ou ses dangers.

Le lecteur pourra, après quelques applications réussies et quelques échecs inévitables, définir les problèmes qu'il se pose d'une façon plus adaptée à l'analyse statistique. Plus précisément, nous distinguons quatre niveaux d'utilisation de la statistique appliquée.

Dans un premier temps en effet, les difficultés rencontrées par l'utilisateur concerneront la façon dont il peut analyser des données : il devra choisir la méthode adéquate, et en interpréter les résultats.

Après quelques expériences, au cours desquelles il s'apercevra par exemple que le traitement statistique d'une question permettant des réponses multiples (par exemple, choisir un ou plusieurs items dans une liste), n'est ni facile ni efficace, il réfléchira sur la rédaction des questionnaires, pour en permettre ou en faciliter l'exploitation.

Une bonne connaissance des méthodes pourra aussi lui donner des idées sur la façon d'aborder les problèmes qu'il rencontre dans son travail : la statistique permettant de répondre à différentes sortes de questions, pourquoi alors ne pas se les poser ?

Enfin, et c'est l'objectif principal de l'auteur, souvent inquiet de la façon dont l'analyse statistique est déviée pour justifier des mesures prises a priori ou des préjugés, il développera une réflexion critique sur les méthodes et une certaine méfiance vis-à-vis des affirmations fondées sur des chiffres auxquelles il se trouvera confronté. Il comprendra alors que l'évaluation quantitative, nécessaire pour synthétiser un ensemble d'observations, donne une image appauvrie de chaque unité statistique.

Les quatre niveaux de ce programme ambitieux sont tous accessibles à des étudiants de formations diverses. L'expérience de l'auteur, dans son enseignement comme dans sa pratique, lui permet d'affirmer que les mathématiciens et les informaticiens ne sont pas a priori favorisés dans cette progression. Beaucoup d'entre eux en effet se contentent de régler les difficultés d'ordre mathématique ou informatique, alors que finalement ce sont les moins intéressantes dans le cadre que nous nous sommes fixé ici. Ils trouveront donc dans cet ouvrage, comme les étudiants en psychologie, en sociologie, en gestion, etc..., une démarche intellectuelle différente dans laquelle il n'est pas indispensable de maîtriser les mathématiques ni l'informatique pour atteindre les objectifs fixés.

La formation à l'analyse des données et à la régression ne se limite pas à la lecture d'un ouvrage sur ces méthodes. La pratique est une source considérable d'apprentissage : c'est elle qui permet de prendre du recul vis-à-vis des méthodes, qui développe l'esprit critique. Tout étudiant devrait

appliquer chacune de ces méthodes sur des données diverses, et en tirer des conclusions à la fois sur les données et sur la méthode.

Pour cela, il est nécessaire d'utiliser un logiciel : ceux qui permettent d'effectuer une analyse factorielle ou une régression sont nombreux. On aura plus de difficultés à trouver un programme d'analyse factorielle d'un tableau de distances, d'analyse discriminante, ou d'analyse des correspondances dans les logiciels anglo-saxons. Il faut parfois faire attention aux résultats numériques et aux graphiques qui ne sont pas toujours construits correctement. Les logiciels ont des conceptions différentes : certains utilisent des " piles ", c'est-à-dire que l'utilisateur définit la suite des traitements qu'il veut appliquer aux données (SAS, CHADOC). D'autres sont très puissants dans la réalisation des graphiques (SPAD, STATLAB) ou ont un mode d'utilisation très conversationnel (PCSM, STATGRAPHICS). Le logiciel S+ est un langage de programmation comportant des commandes statistiques très puissantes. Le logiciel le plus répandu dans les entreprises est semble-t-il SAS qui donne facilement accès aux bases de données. Il existe des logiciels peu chers comme STAT-ITCF (dont l'auteur a rédigé jadis les programmes d'analyse factorielle). Le logiciel STATPC , qui fonctionne sur les micro-ordinateurs de type P.C. sous MS-DOS est disponible en freeware auprès de l'auteur (e-mail : foucart@mathrs.univ-poitiers.fr). Tous les calculs figurant dans l'ouvrage ont été réalisés avec STATPC.

Avec cet ouvrage et un logiciel d'analyse de données, les étudiants sont en mesure de comprendre les méthodes statistiques et de les appliquer correctement sur des données diverses. Ils n'ont plus qu'à travailler. L'auteur souhaite que cet ouvrage les aide dans leurs études et les accompagne dans leur pratique professionnelle.

Poitiers, le 10 Juin 1997

Thierry Foucart

Remerciements : l'auteur remercie Sylvie Trichard, Jean Cougnard, Farid Beninel et Patrick Plassars de leurs remarques pertinentes.

N.B. : le logiciel STATPC et cet ouvrage sur disquette (format WORD6 PC) sont disponibles auprès de l'auteur (e-mail : foucart@mathrs.univ-poitiers.fr).